Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie
ISSN: (Online) 2222-4173, (Print) 0254-3486

’n Oorsig van grootdata- en datawetenskaponderrig aan Suid-Afrikaanse universiteite

Author:
Eduan Kotzé[1]

Affiliation:
[1]Department of Computer Science and Informatics, University of the Free State, South Africa

Corresponding author:
Eduan Kotzé,
kotzeje@ufs.ac.za

Dates:
Received: 19 Apr. 2016
Accepted: 20 June 2016
Published: 03 Nov. 2016

How to cite this article:
Kotzé, E., 2016, ‘’n Oorsig van grootdata- en datawetenskaponderrig aan Suid-Afrikaanse universiteite’, *Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie* 35(1), a1387. http://dx.doi.org/10.4102/satnt.v35i1.1387

Mens en rekenaar genereer data elektronies teen ’n astronomiese spoed en in so ’n mate dat die samelewing tans kognitiewe uitdagings ervaar om al hierdie data sinvol te ontleed en te gebruik. Grootdatafirmas, soos Google en Facebook, het hierdie probleem reeds ’n paar jaar gelede geïdentifiseer en is tans gedurig besig om nuwe tegnologie te ontwikkel of bestaande tegnologie te verbeter, ten einde die kognitiewe ontleedproses van hierdie grootdatastelle te vergemaklik. Die doel van dié artikel is om ’n bydrae te lewer tot ons teoretiese begrip van die rol wat grootdata gaan speel om nuwe opleidingsgeleenthede aan Suid-Afrikaanse universiteite te skep. Die artikel ondersoek ontluikende literatuur oor die eienskappe en hoofkomponente van grootdata, tesame met die Hadoop-programstapel as voorbeeld van grootdatategnologie. Aangesien die snelle ontwikkeling van grootdatategnologie ’n paradigmaverskuiwing teweeg bring wat betref menslike hulpbronne wat hierdie datastelle gaan ontleed, stel die studie ondersoek in na die stand van grootdata-onderrig aan Suid-Afrikaanse universiteite. Die artikel voorsien ook ’n oorsig oor moontlike grootdatabronne aan Suid-Afrikaanse universiteite, sowel as relevante grootdatavaardighede waaroor die datawetenskaplike noodwendig moet beskik. Die studie ondersoek verder ook bestaande akademiese programme in Suid-Afrika, wat gefokus is op die onderrig van gevorderde databasisstelsels. Die studie het bevind dat grootdata-onderrig wel op nagraadse vlak aan Suid-Afrikaanse universiteite plaasvind, maar op ’n baie beperkte skaal. Dié artikel dra by tot nuwe kennis en stel belangrike teoretiese komponente voor wat ingesluit kan word by bestaande akademiese programme. Daar is egter meer navorsing nodig om hierdie programme uit te brei en in ’n groeiende vraag aan datawetenskaplikes met grootdatavaardighede te voorsien.

An overview of big data and data science education at South African universities. Man and machine are generating data electronically at an astronomical speed and in such a way that society is experiencing cognitive challenges to analyse this data meaningfully. Big data firms, such as Google and Facebook, identified this problem several years ago and are continuously developing new technologies or improving existing technologies in order to facilitate the cognitive analysis process of these large data sets. The purpose of this article is to contribute to our theoretical understanding of the role that big data might play in creating new training opportunities for South African universities. The article investigates emerging literature on the characteristics and main components of big data, together with the Hadoop application stack as an example of big data technology. Due to the rapid development of big data technology, a paradigm shift of human resources is required to analyse these data sets; therefore, this study examines the state of big data teaching at South African universities. This article also provides an overview of possible big data sources for South African universities, as well as relevant big data skills that data scientists need. The study also investigates existing academic programs in South Africa, where the focus is on teaching advanced database systems. The study found that big data and data science topics are introduced to students on a postgraduate level, but that the scope is very limited. This article contributes by proposing important theoretical topics that could be introduced as part of the existing academic programs. More research is required, however, to expand these programs in order to meet the growing demand for data scientists with big data skills.

Inleiding

Volgens IBM word daar daagliks ongeveer 2.5 kwintiljoen *(quintillion)* datagrepe deur ’n verskeidenheid bronne gegenereer wat byvoorbeeld klimaatsensors, kommersiële transaksies en ander digitale media insluit (IBM n.d.). Om hierdie getal in perspektief te plaas word daar dikwels verwys daarna dat 90% van die wêreld se data in die laaste paar jaar genereer is. Verder is 90%

van hierdie data egter in 'n ongestruktureerde formaat. Vir die mens is dit baie moeilik om ongestruktureerde data te sintetiseer en te kontekstualiseer ten einde enige waarde daaruit te onttrek. Yin en Kaynak (2015) voer aan dat hierdie enorme hoeveelheid data, beter bekend as grootdata, 'n belangrike rol gaan speel in die toekoms en nie geïgnoreer kan word nie. Die outeurs skets 'n toekomsprentjie van 'n vierde industriële revolusie met 'slim' fabrieke, waar masjiene en bronne deur sosiale netwerke gaan kommunikeer. Hierdie fabrieke se einddoel is om grootdatabronne te gebruik om 'slim' produkte te lewer wat sal weet hoeveel reeds vervaardig is en hoeveel deur die mark benodig word. Davenport en Patil (2012) beklemtoon dat datawetenskaplikes 'n baie belangrike rol gaan speel om grootdatastelle te bou en te ontleed. Daar is dus tans wêreldwyd 'n fokusverskuiwing by akademiese instellings om datawetenskaplikes op te lei (Wixom *et al.* 2014), en lande wat daarin gaan slaag, sal sekerlik 'n voorsprong hê in die vierde industriële revolusie.

Die artikel ondersoek die stand van grootdata- en datawetenskaponderrig aan Suid-Afrikaanse universiteite. Die artikel begin met 'n bespreking van die eienskappe, gevolg deur die fases van 'n grootdatastelsel. In die daaropeenvolgende afdeling word enkele maar toepaslike grootdata-uitdagings bespreek as potensiële navorsingsgeleenthede. Daarna word gefokus op grootdata in hoër onderwys en bespreek enkele tipes grootdatabronne wat by universiteite teenwoordig is. Dié afdeling lewer ook verslag oor die stand van grootdata- en datawetenskaponderrig aan Suid-Afrikaanse universiteite en vorm die empiriese gedeelte van die artikel. Ten slotte word opsommende kommentare en voorstelle vir toekomstige navorsing bespreek.

Grootdata

Die hoofdoel van hierdie afdeling is om 'n oorsig te gee oor grootdata. Grootdata-eienskappe, grootdatastelsels en die Hadoop-programmatuurstapel word breedvoerig bespreek.

Grootdata-eienskappe

Grootdata verwys na 'n groot versameling data wat óf ongestruktureerd is, óf so vinnig groei dat dit te groot is vir tradisionele dataverwerkingstelsels en dus nuwer tegnologie vereis (Vaisman & Zimanyi 2014; Provost & Fawcett 2013). Wat grootdata so enorm maak, is herhaalde inskrywings oor 'n tydperk, byvoorbeeld, 'n weblog skryf miljoene besoeke per dag op 'n paar webblaaie; of 'n selfoondatabasis stoor die tyd en plek van al die handstelle elke 15 sekondes (Jacobs 2009). Die datastel moet ook so groot wees dat tradisionele inligtingstegnologie nie die data binne 'n aanvaarbare tyd kan verwerk nie. Hu *et al.* (2014) vergelyk grootdata met tradisionele data en wys die verskille uit ten opsigte van volume, skeppingspoed, struktuur, databron, databerging, data-integrasie en toegang. Grootdata se volumes word in tera- of petagrepe gemeet, dit word konstant opgedateer, is semigestruktureerd of ongestruktureerd, ten volle verspreid, moeilik om te integreer en word meestal in nierelasionele databasisse geberg (Hu *et al.* 2014).

Volgens Gartner (2016) word grootdata met drie hoofeienskappe geassosieer, naamlik groot hoeveelhede data (*volume*), wye verskeidenheid datatipes (*variety*) en hoë snelheid (*velocity*). Ander eienskappe van grootdata sluit in laedigtheidwaarde (*value*), betroubaarheid (*veracity*), verifikasie (*verification*), en validasie (*validation*) (Berman 2013; Wamba *et al*. 2015; Jagadish *et al.* 2014; Yin & Kaynak 2015; Assunção *et al.* 2015). Waarde verwys na die vermoë om insig uit grootdata te geneer tot voordeel van 'n organisasie (Daniel & Butson 2014). Betroubaarheid verwys na die teenwoordigheid van vuilheid en onakkuraatheid wat deur geruis en abnormaliteite veroorsaak word (Daniel & Butson 2014). Dataverifikasie verwys na 'n stel prosesse wat gevolg moet word om te verseker dat 'n grootdatastel konformeer na 'n stel spesifikasies voordat dit deur analiste gebruik word (Berman 2013). Die laaste eienskap, datageldigheid, is 'n eienskap waar grootdata op 'n gepaste wyse aangewend word sodat dit in ooreenstemming is met die oorspronklike doelwit van 'n grootdata-projek (Berman 2013).

Grootdatastelsel

Bogenoemde eienskappe van grootdata bring 'n nuwe paradigma na vore met betrekking tot hoe die data versamel, geberg, verwerk, ontleed en gebruik moet word. Hierdie paradigma word in die gedrukte media gewoonlik net met tegnologie geassosieer, en dikwels word daar nagelaat om te sê dat grootdata ook twee ander kritieke elemente vereis, naamlik mense en 'n proses (W. Kim, Jeong, & C. Kim 2014). Die volgende afdeling snoer die elemente saam en bespreek die vier aaneenlopende fases van 'n grootdatastelsel. Hierdie fases sluit in databronne, dataverkryging, databerging en data-ontleding (Yin & Kaynak 2015; Hu *et al*. 2014; Chen *et al.* 2014).

Datagenerasie en databronne

Van die belangrikste primêre bronne van grootdata sluit in databanke, webblaaie, sosiale media, sensors en mobiele data (Hu *et al.* 2014). Hierdie bronne word hoofsaaklik deur sakeondernemings, rekenaarnetwerke wat die Internet, mobiele netwerke en Internet of Things insluit, sowel as deur wetenskaplike toepassings genereer (Hu *et al*. 2014). Ten einde as 'n grootdatabron gesien te word, moet groot hoeveelhede data teen 'n baie hoë spoed gegenereer word. Grootdatabronne kom in verskeie formate voor en kan gestruktureerd, semigestruktureerd of ongestruktureerd wees (Hu *et al.* 2014). Gestruktureerde data verwys na data wat in entiteite georganiseer is, soos XML-dokumente of databasistabelle wat in 'n relasionele databasisstelsel gevind word (White 2015). Semigestruktureerde data het wel 'n skema, maar is minder georganiseerd en gestruktureerd (B. G. Kim, Trimi & Chung 2014). 'n Spreivel (*spreadsheet*) is 'n voorbeeld van 'n semigestruktureerde bron (White 2015). Ongestruktureerde data beskik nie oor 'n interne struktuur nie en sluit in teksdata, klankdata, videodata, beelddata en internetdata soos klikstrome (*click stream*) en loglêers (White 2015; Minelli, Chambers & Dhiraj 2013: 10). Om sake verder te kompliseer, kan grootdata ook arriveer en verwerking

benodig teen verskillende snelhede, wat insluit in bondels, nabytyds (*near-time*), intyds (*real-time*) of in strome (Assunção *et al.* 2015). Die verspreide MapReduce en Hadoop-ekosisteem is die ideale omgewing om databronne te akkommodeer wat in verskeie formate en snelhede voorkom, en word breedvoerig in die Hadoop-programmatuurstapel bespreek.

Dataverkryging

Die verkryging van grootdatastelle behels die proses om data bymekaar te maak en sluit in data-insameling, data-oordrag en data-voorverwerking (Chen *et al*. 2014:23). Die drie mees algemene maniere om grootdata te versamel sluit in loglêerverkryging, sensordataverkryging en netwerkdataverkryging (Hu *et al.* 2014). Loglêerverkryging is die mees algemene manier om grootdata te verkry, aangesien dit deur databronstelsels en webbedieners gegenereer word (Chen *et al.* 2014:24). Sensordata kom ook volop voor en sluit in enige data wat deur elektroniese sensors gegenereer word (Chen *et al.* 2014:24). Netwerkdata, soos webblaaie, word verkry deur 'n kombinasie van 'n webkruiper (*web crawler*) en 'n woordsegmentasiestelsel (Chen *et al.* 2014:24). Sodra die data versamel is, word dit na 'n datasentrum oorgeplaas waar analitiese toepassings ondersteun word. Die proses word gevolg deur 'n datasentrumoordragproses waar die data binne die datasentrum oorgedra word vir plasingoordeel (Hu *et al.* 2014). Vanweë die groot verskeidenheid databronne, gaan die versamelde data geruis en oorbodigheid bevat en is dit nie sinvol om onbruikbare data te stoor nie (Chen *et al.* 2014:27). Die data word dus deur 'n datavoorverwerkingsfase gehanteer waar dit geïntegreer, skoongemaak en oorbodighede verwyder word ten einde stoorspasie te verminder en ontledingsakkuraatheid te verhoog (Hu *et al*. 2014; Chen *et al.* 2014:27).

Databerging

Sodra die data versamel en verkry is, moet 'n grootdatastelsel dit gereed kry vir berging, ontleding en waardeontrekking (Hu *et al.* 2014). Dit alles bied interessante uitdagings vir 'n databergingstelsel, aangesien gestruktureerde data herhaaldelik deur 'n relasionele databasisbestuurstelsel (RDBS) gehanteer word, terwyl semigestruktureerde of ongestruktureerde data meer op 'n ad hoc-basis, of selfs net 'n enkele keer gehanteer word (Hu *et al.* 2014). Vir hierdie doeleindes moet buigsame databergingsinfrastruktuur in plek wees en die databergingstelsel moet ook 'n skaalbare toegang-koppelvlak voorsien vir navrae na die data. 'n Tipiese databergingstelsel bestaan uit apparatuurinfrastruktuur en databestuur.

Apparatuurinfrastruktuur, byvoorbeeld DAS (*Direct Attached Storage*), NAS (*Network Attached Storage*) en SAN (*Storage Area Network*), is verantwoordelik daarvoor om die versamelde inligting fisiek te stoor (Hu *et al.* 2014). Databestuur is weer verantwoordelik daarvoor om die inligting te organiseer sodat dit effektief verwerk kan word. Grootdatabestuur bestaan uit drie vlakke, naamlik lêerstelsels, databasistegnologie en programmeringsmodelle. Die Google-lêerstelsel (*Google File System* of GFS), Cosmos van Microsoft en Haystack van Facebook is bekende voorbeelde van verspreide lêerstelsels (Hu *et al.* 2014). Van al hierdie lêerstelsels is GFS die gewildste, aangesien dit 'n brokkiegebaseerde (*chuck-based*), verspreide lêerstelsel is wat fouttoleransie deur middel van datapartisie en replisering ondersteun (Lee *et al*. 2011). Aangesien GFS 'n gedeeldelêerstelsel is, funksioneer dit die effektiefste op 'n rekenaartros (*cluster*) wat bestaan uit 'n stel gekoppelde rekenaars wat saamwerk as 'n enkele sisteem.

Ten opsigte van databestuur en databasistegnologie, word daar algemeen aanvaar dat 'n tradisionele RDBS nie die verskeidenheid en grootte van grootdatastelle kan hanteer nie (Hu *et al.* 2014; Krishnan 2013: 86). 'n RDBS ondersteun net gestruktureerde data (Hu *et al*. 2014) vanweë die relasionele argitektuur en ACID-eienskappe (*Atomicity, Consistency, Isolation and Durability*) van 'n RDBS (Krishnan 2013: 86). 'n RDBS kan egter ook slegs opskaal deur middel van peperduur apparatuur en nie deur kommoditeitsapparatuur, soos by 'n rekenaartros wat parallel werk nie (Hu *et al.* 2014). Beide hierdie faktore (struktuur en skaalbaarheid), maak 'n RDBS dus onvanpas om die groeiende volumes data van nuwerwetse toepassings, soos sensornetwerke en ruilhandelplatforms, te bestuur (Krishnan 2013:86; Hu *et al.* 2014).

NoSQL- (Not Only SQL) databasisstelsels word tans voorgehou as die oplossing om hierdie groot volumes, multistrukturele data te verwerk (Sharda *et al.* 2014). NoSQL-databasisse kan min of meer in vier hoofkategorieë verdeel word: sleutelwaarde, kolomgeoriënteerd, dokumentdatabasisse en grafiekdatabasisse (Hu *et al*. 2014). 'n Sleutelwaardedatabasis het 'n eenvoudige datamodel waar data as 'n sleutelwaardepaar geberg word en elke sleutel uniek is (Hu *et al.* 2014). Bekende voorbeelde van hierdie tipe databasis sluit in Voldemort, wat deur LinkedIn.com ontwikkel is en Dynamo, wat deur Amazon se e-besigheidsplatform gebruik word om datagedrewe aanbevelings vanuit grootdata te verkry (Chen *et al.* 2014:38; Provost & Fawcett 2013). Kolomgeoriënteerde databasisse berg en verwerk data deur kolomme en nie deur rye soos in die geval van relasionele databassisse nie (Hu *et al*. 2014). Bekende voorbeelde van hierdie tipe databasis is Google se BigTable, Cassandra en HBase (Chen *et al.* 2014;41). HBase is 'n oopbron-kloon van BigTable en is deel van die Apache Hadoop en MapReduce- raamwerk (Chen *et al.* 2014:41) en word in Hadoop-programmatuurstapel bespreek. Dokumentdatabassise ondersteun meer komplekse sleutelwaardestore waar die data as 'n dokument geberg word, en in JSON-formaat voorgestel word (Hu *et al.* 2014). Bekende voorbeelde van hierdie tipe databasis is MongoDB, SimpleDB, CouchDB , sowel as Riak (Krishnan 2013:86). Grafiekdatabasisse is gebaseer op grafiekteorie en ondersteun skaalbaarheid oor 'n tros van rekenaars (Krishnan 2013:86). Dit word ook gesien as een van die mees komplekse NoSQL-databasistipes en het ontwikkel vanweë die geweldige toename in data uit sosiale media (Krishnan 2013:97).

'n Voorbeeld van hierdie tipe databasis is Neo4J (Krishnan 2013:86).

Data-ontleding en waardeontrekking

Datadatavisualisering, statistiese ontleding en data-ontginning word met groot sukses in verskeie grootdata-ontledingtoepassings gebruik (Hu *et al.* 2014). Hierdie toepassings sluit in gestruktureerde data-ontleding, teksontginning, webontginning, multimedia-ontleding en mobiele-ontleding (Hu *et al.* 2014). Gestruktureerde data-ontleding, soos data-ontginning en statistiese ontleding, word meestal op sake- en wetenskaplike eksperimentdata gebruik waar die rekords gestruktureerd is (Hu *et al.* 2014). Teksontginning van ongestruktureerde bronne, soos e-posse, loglêers en teksgedeeltes van webblaaie, word met groot sukses uitgevoer om besigheidsinligting te identifiseer en te onttrek (Sabherwal & Becerra-Fernandez 2011:87). Die belangrikste teksontginningstegnieke sluit in natuurlike taalverwerking (*Natural Language Processing* of NLP), statistiese en ko-voorkomsontleding, neurale netwerke en visualisering (Sabherwal & Becerra-Fernandez 2011:87). Waar teksontginning die inhoud van ongestruktureerde bronne ontgin, fokus webontginning op die deursoek van aanlynbronne, soos webblaaie. Webontginning word hoofsaaklik op drie wyses uitgevoer, insluitend inhoud-, struktuur- en gebruikontginning (Hu *et al*. 2014; Sabherwal & Becerra-Fernandez 2011). Die databron is die teksdata van 'n webbladdokument en die drie wyses maak staat op tegnieke soos inligtingonttrekking, kunsmatige intelligensie en tekskategorisering om nuwe kennis te ontdek (Sabherwal & Becerra-Fernandez 2011). Multimedia-ontleding maak gebruik van databronne soos beeld, klank en video om interessante kennis te onttrek en dan die semantiek te probeer verstaan soos dit vasgevang is in die multimediadata (Hu *et al.* 2014). Van die belangrikste navorsing sluit in multimedia-opsomming, multimedia-annotasie, multimedia-indeksering en -onttrekking, multimedia-aanbeveling, en multimedia-gebeurtenisopsporing. Met die toename in die getal mobiele rekenaars, het mobiel-ontleding belangriker geword. Van die belangrikste ontwikkelings sluit in radiofrekwensie-identifikasie (*Radio Frequency Identification of RFID*) en koordlose sensors. RFID word al hoe meer in voorraadbestuur en logistieke aktiwiteite gebruik, terwyl mobiele sensors, soos die intydse monitering van 'n persoon se gesondheid, groot verandering in mediese sorg teweeg gebring het (Hu *et al.* 2014).

Hadoop-programmatuurstapel

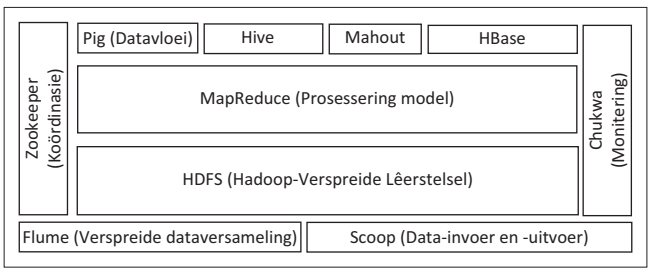
Die Apache Hadoop-sagtewarebiblioteek is 'n raamwerk wat uit verskeie oopbronprogrammatuur bestaan en implementeer hoofsaaklik Google se MapReduce-infrastruktuur (Lee *et al.* 2011). Die raamwerk voorsien 'n verspreide dataverwerkingsomgewing deur van 'n rekenaartros gebruik te maak (Hu *et al.* 2014). Hadoop se hoofdoel is dus om groot volumes verspreide gestruktureerde, semigestruktureerde en ongestruktureerde data deur middel van programmeringmodelle te verwerk, te stoor en te ontleed (Vaisman & Zimanyi 2014:508; Sharda *et al.* 2014:588). Dit is egter beter om na 'n Hadoop-ekosisteem (Eckerson 2011) te verwys, eerder as net na 'n enkele komponent, wanneer grootdata bespreek word.

Die Hadoop-ekosisteem bestaan uit drie hoofkomponente, naamlik 'n verspreide-lêerstelsel of datastoorvlak wat *Hadoop Distributed File System* (HDFS) genoem word, 'n hulpbronskeduleerder wat *Yet Another Resource Negotiator* (YARN) genoem word, en 'n dataverwerkingsvlak wat die verspreide MapReduce-verwerkingsraamwerk insluit (Lee *et al.* 2011; White 2015; Vavilapalli *et al.* 2013). 'n Diagrammatiese voorstelling (Figuur 1) van 'n tipiese Hadoop-ekosisteem, soos uit Hu *et al.* (2014), gaan nou gebruik word om die belangrikste oopbronprogrammatuur in die ekosisteem te bespreek.

MapReduce en *Yet Another Resource Negotiator*

MapReduce is oorspronklik deur Google ontwikkel om websoektogte op webdokumente deur middel van 'n groot aantal rekenaars te verrig (Vaisman & Zimanyi 2014). Deesdae word MapReduce gebruik as 'n programmeringsmodel om grootdatastelle te genereer en te verwerk (White 2015:19; Sharda *et al.* 2014:587). Van al die belangrikste programmeringsmodelle, wat onder meer Dryad, Pregel, GrahpLab, S4 en Storm insluit, is MapReduce die gewildste (Hu *et al*. 2014). MapReduce is nie 'n programmeringstaal nie en is ontwerp om deur programmeerders eerder as die sakewêreld gebruik te word (Sharda *et al.* 2014:587). Verder gebruik MapReduce ook GFS as onderliggende bergplek, wat dit 'n geskikte omgewing maak om 'n hele datastel as 'n bondel te ontleed. Dit is in teenstelling met 'n RDBS, wat eerder geskik is om navrae en opdaterings van klein hoeveelhede data te hanteer (White 2015:8). Om daardie rede behoort MapReduce gesien te word as bykomende tegnologie tot 'n RDBS en nie as 'n plaasvervanger nie (Russom 2015).

MapReduce kan oor verskeie dataformate geïmplementeer word en werk volgens die beginsel waar 'n taak in kleiner dele verdeel word, en dan in parallel verwerk word oor 'n versameling identiese masjiene, soos 'n rekenaartros (Vaisman & Zimanyi 2014:508). Volgens Chen (2014:16) bestaan die MapReduce-raamwerk uit een JobTracker-node en verskeie TaskTrackers-nodes. Hierdie nodes werk saam en stel MapReduce-programme in staat om nuttige data vanuit 'n grootdatastel te onttrek. 'n Tipiese MapReduce-program



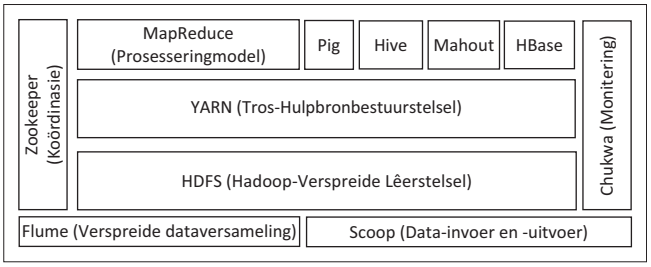
Bron: Hu, H., Wen, Y., Chua, T. & Li, X., 2014, 'Toward scalable systems for big data analytics: A technology tutorial', *Access, IEEE*, vol. 2, pp.652–687

FIGUUR 1: Hadoop 1.0-programmatuurstapel.

bestaan weer uit twee take, naamlik Kaart (*Map*) en Verminder (*Reduce*) (Vaisman & Zimanyi 2014:508). Die JobTracker-node is verantwoordelik daarvoor om 'n taak te versprei en te skeduleer, terwyl die TaskTracker-node gebruik word om die Kaart- of Vermindertake wat deur die JobTracker-node versprei is, te ontvang. Die TaskTracker-node voer die take in parallel op 'n trosomgewing uit en stuur dan die terugvoerstatus na die JobTracker-node (Chen *et al.* 2014:16). Die eindproduk van hierdie proses is 'n lêer wat kan dien as invoer vir 'n datapakhuis of ontleed kan word deur sake-intelligensiegereedskap soos Tableau, Cognos of Microstrategy. YARN, wat Hadoop se tros-hulpbronbestuurstelsel is, word sedert Hadoop weergawe 2.0 gebruik om die werkverrigting van MapReduce-implementerings te verbeter (Vavilapalli *et al.* 2013; Hadoop n.d.). In Hadoop 1.0 was MapReduce verantwoordelik vir beide tros-hulpbronbestuur sowel as dataverwerking. In YARN, wat beskou word as MapReduce 2, hanteer MapReduce net dataverwerking en is YARN nou verantwoordelik vir tros-hulpbronbestuur. Hierdie verdeling van take beteken die nuwe ekosisteem kan nie net beter skaal nie, maar kan ook meer nodusse akkommodeer en is dus 'n verbetering op die oorspronklike Hadoop 1.0-ekosisteem (White 2015:79). Sien Figuur 2 vir 'n diagrammatiese voorstelling van die Hadoop 2.0-ekosisteem.

Hadoop-verspreidelêerstelsel

Die Hadoop-verspreidelêerstelsel se hoofverantwoordelikheid is om grootdatastelle op te deel tussen rekenaars, of nodes in 'n HDFS-tros (White 2015). 'n HDFS-tros bestaan gewoonlik uit 'n enkele NaamNode wat die lêerstelsel se metadata bestuur, en 'n versameling DataNodes wat die werklike data stoor (Hu *et al.* 2014). Aangesien HDFS lêergebaseer is, benodig dit nie 'n datamodel soos in die geval van 'n RDBS om die data te stoor of te verwerk nie en kan dit data van enige struktuur stoor. Sodra 'n lêer op HDFS opgelaai word, word die lêer opgedeel in blokke, die blokke word dan versprei tussen rekenaars binne die HDFS-tros en laastens, word die blokke gedupliseer sodat veelvoudige kopieë van elke blok op verskeie rekenaars binne die HDFS-tros geberg word (White 2015). HDFS tesame met Hadoop voorsien dus die perfekte omgewing om semigestruktureerde en ongestruktureerde data te stoor en deur middel van MapReduce-programme, wat meestal in parallel uitgevoer word, te verwerk (White 2015; Watson 2014). Uit bogenoemde bespreking kan die gevolgtrekking gemaak word dat Hadoop en MapReduce altyd saam moet funksioneer. Dit is egter misleidend, aangesien hulle mekaar nie nodig het om selfstandig te kan funksioneer nie. Ontwikkelaars by Google het MapReduce ontwikkel reeds voor die bestaan van HDFS, en sekere variasies van MapReduce werk met verskeie ander soorte stoortegnologie en selfs 'n ander RDBS (Sharda *et al.* 2014:592).



Bron: Hu, H., Wen, Y., Chua, T. & Li, X., 2014, 'Toward scalable systems for big data analytics: A technology tutorial', *Access, IEEE*, vol. 2, pp.652–687; Hadoop, n.d., 'Hadoop documentation and open source release', viewed on 08 March 2016, from: http://hadoop.apache.org/

FIGUUR 2: Hadoop 2.0-programmatuurstapel.

ZooKeeper en *Chukwa*

Beide Zookeeper en Chukwa word gebruik om verspreide toepassings wat op Hadoop uitgevoer word te moniteer en te bestuur (Hu *et al.* 2014). Zookeeper is oorspronklik deur Yahoo! ontwikkel om dit makliker vir toepassings te maak om toegang tot konfigurasie-inligting te verkry, maar het intussen so gegroei dat dit toepassings oor verspreide rekenaartrosse kan help koördineer (Warden 2011:10). Zookeeper word deesdae beskou as 'n gespesialiseerde sleutelwaarde-stoor met 'n koppelvlak wat soos 'n lêerstelsel funksioneer (Warden 2011). Chukwa is 'n Hadoop-subprojek wat dien as dataversamelingstelsel (Warden 2011) waarmee groot verspreide stelsels gemoniteer en bestuur word (Krishnan 2013). Chukwa is op HDFS en die MapReduce-programmeringsmodel gebou en beskik oor buigsame en kragtige gereedskap wat resultate vertoon, moniteer en ontleed (Hadoop n.d.). Chukwa word reeds met groot sukses deur verskeie organisasies gebruik om logdata te versamel en met MapReduce te ontleed (Rabkin & Katz 2010).

Hbase

Hbase is 'n verspreide kolomdatabasis wat Hadoop in staat stel om groot tabelle te stoor (Watson 2014). Hbase is gemodelleer volgens Google se BigTable-datastoorstelsel (Chang *et al*. 2006) en lewer saam met Hadoop, 'n foutverdraagsame omgewing waar grootdatatabelle gemanipuleer en geberg kan word (Loshin 2013:67). Hbase vervang GFS met HDFS en skryf inhoud gereeld vanuit interne geheue na lêers (Chen *et al.* 2014:41). Aangesien HBase 'n NoSQL-databasis is, ondersteun dit nie SQL-navrae nie, maar wel verkry-, plaas-, skandeer- en uitveeopdragte (Loshin 2013:67). Hbase ondersteun ook ingeheue-uitvoering, wat 'n eienskap is van die nuwe generasie ingeheue-databasisstelsel (*in-memory database system*) (Vaisman & Zimanyi 2014). Voorbeelde van bekende internetfirmas wat HBase as databasisstelsel gebruik, is eBay.com en Facebook.com (Sharda *et al*. 2014:590).

Hive en *Pig*

Vir gebruikers wat nie vertroud is met die skryf van MapReduce-kode nie, is Hadoop nie maklik om te gebruik nie, aangesien MapReduce-programme meestal in Java, Python, C of Perl geskryf word (Watson 2014). Om daardie rede is Hive en Pig ontwikkel, en word programmeerders toegelaat om op 'n hoër abstraksievlak te werk, in plaas van Java of ander laevlaktale wat deur Hadoop ondersteun word (Hu *et al.* 2014; Vaisman & Zimanyi 2014). Beide Hive- en Pig-programme word intern omgeskakel na MapReduce-take, sonder dat die programmeerder die oorskakeling hoef te behartig (Vaisman & Zimanyi 2014).

Hive is deur Facebook ontwikkel om die konsepte van tabelle, kolomme en SQL (soos uit die relasionele databasiswêreld) na die Hadoop-ekosisteem te bring (Vaisman & Zimanyi 2014). Hive word gesien as die datapakhuisstelsel vir Hadoop en maak ook voorsiening vir ad hoc-navrae, data-opsomming, asook die onttrekking, transformasie en laai van data (*extraction, transformation and loading* [ETL]) na 'n verskeidenheid dataformate (Loshin 2013:68). Hive laat gebruikers toe om grootdata in tabelle en partisies te organiseer en voorsien ook HiveQL, 'n SQL-dialek waarmee gebruikers navrae kan skryf (Vaisman & Zimanyi 2014). Hive is ook verantwoordelik om 'n HiveQL-navraag na 'n gerigte asikliese grafiek (*directed acyclic graph* [DAG]) van MapReduce-take te vertaal (Lee *et al.* 2011). Aangesien die Hive op Hadoop gebou is, gee dit programmeerders indirek toegang tot die MapReduce-model. Dit stel hulle in staat om pasgemaakte kaart- en verminderingsfunksies van MapReduce te ontwikkel wat met HiveQL-navrae integreer (Loshin 2013:68). Een van die tekortkominge van Hive is die onvermoë om intydse navrae of ry-opdaterings te hanteer (Leverenz 2015). Hive se beste toepassing is dus eerder ontleding en nie gekoppelde transaksieverwerking (*online transaction processing* [OLTP]) of bondelverwerking nie (Celko 2014:60).

Pig bestaan uit datavloeitaal op hoë vlak genaamd Pig Latin en 'n gepaardgaande uitvoeringsraamwerk (Lee *et al.* 2011). Pig Latin is deur Yahoo! ontwikkel as 'n hoëvlak-programmeringstaal om navrae te hanteer op HDFS-data en word as deel van die Hadoop-ekosisteem beskou (Vaisman & Zimanyi 2014; Watson 2014). Aangesien Pig 'n prosedurele programmeringstaal is wat hoëvlak-abstraksies vir MapReduce verskaf (Intel 2013), kan opdragte op drie maniere uitgevoer word, naamlik geskrewe teks (*script*), bevellynopdrag en om Pig-opdragte te roep wat in Javabronkode ingebed is (Vaisman & Zimanyi 2014).

Mahout

Mahout is 'n oopbron-biblioteek van masjienleer en data-ontginningalgoritmes wat onder meer groepering, kategorisering, samewerkende filter, teksontginning en gereelde patroonontginning insluit (Watson 2014; Loshin 2013:70–71). Mahout implementeer hierdie algoritmes op Hadoop deur weereens van die MapReduce-model (Sharda *et al.* 2014:590) en bondelverwerking (Chen *et al.* 2014:17) gebruik te maak. Van die belangrikste voordele wat Mahout inhou, is dat dit ontwerp is om met minimale opstelling op Hadoop ontplooi te kan word en ook baie effektief kan skaal (Celko 2014:240).

Flume en *Scoop*

Apache Flume en Scoop is gereedskap wat ontwikkel is om te help met data-integrasie. Flume is 'n verspreide stelsel wat groot hoeveelhede data van verskeie bronne versamel, saamvoeg en dan oordra na 'n gesentraliseerde datastoor of HDFS (Hu *et al*. 2014; Intel 2013). Voorbeelde sluit in loglêers vanaf toepassingsbedieners of argiefloglêers van ander stelsels (Intel 2013). Scoop, aan die ander kant, hanteer die invoer en uitvoer van data tussen relasionele databasisse en Hadoop (Hu *et al*. 2014). Scoop kan data van byvoorbeeld 'n MySQL of Oracle-databasis na HDFS stuur, 'n MapReduce-taak uitvoer, en die MapReduce-resultate van HDFS dan weer as invoer na 'n relasionele databasis stuur (Intel 2013; Celko 2014:60). Scoop speel dus 'n belangrike rol om data vanaf 'n relasionele databasis na Hadoop in te voer en vergemaklik daarmee die taak van data-analiste.

Apache Spark

Apache Spark is 'n oopbron-trosrekenaarraamwerk vir grootskaalse dataverwerking en kan as alternatief tot Hadoop/MapReduce gebruik word (White 2015:549; Rahman & Iverson 2015). Apache Spark maak gebruik van interaktiewe bondelverwerking, in teenstelling met Hadoop/MapReduce se tradisionele bondelverwerking (Zaharia *et al*. 2010). Verder beskik Apache Spark oor die vermoë om groot datastelle in-geheue te stoor en tussen verskillende verwerkingsopdragte te verwerk (Apache Spark n.d.). Hierdie vermoë maak dit baie aantreklik vir intydse dataverwerking en interaktiewe toepassings omdat die datastelle nie afhanklik is van skyfverwerking, soos in die geval van Hadoop/MapReduce nie (White 2015:550; Landset *et al.* 2015). Apache Spark is ook vinniger, buigsamer en makliker om te programmeer as Hadoop/MapReduce, aangesien toepassings met hoëvlak-programmeringstale soos Python, R en Scala ontwikkel kan word (Landset *et al.* 2015). Die Apache Spark-ekosisteem bestaan uit Spark SQL vir gestruktureerde data, GraphX vir grafiekverwerking, MLLib vir masjienleer en Spark Streaming vir datastroming (Databricks 2016). Dit is belangrik om te noem dat die ekosisteem versoenbaar is met Hadoop-data (HDFS, Hive en HBase) en dat YARN gebruik kan word om Apache Spark-opdragte uit te voer op 'n Hadoop-trosrekenaar (Apache Spark n.d.; Vavilapalli *et al*. 2013).

Kommersiële grootdatastelsels

Nie alle grootdatastelsels is as oopbronprogrammatuur soos Apache Hadoop of Apache Spark beskikbaar nie. Verskeie databasisverskaffers het ook grootdatastelsels beskikbaar gestel, gewoonlik ten duurste. Een so 'n voorbeeld is SAP HANA, 'n ingeheue-databasisplatform vir intydse ontledings en toepassings (Minelli *et al.* 2013:.8). Ander voorbeelde van kommersiële grootdatastelsels wat hulle eie weergawe van Hadoop geskep het, sluit in Cloudera Hadoop, Hortonworks Hadoop, EMC Hadoop, Microsoft Hadoop, Intel Hadoop en MapR (B. G. Kim *et al.* 2014; Davenport 2014). Van die belangrikste ontledings en toepassings waar kommersiële grootdatastelsels tans gebruik word, sluit in geteikende kliëntbemarking, sosialemedia-ontledings, bedrogopsporing-algoritmes en webwerf-aanbevelingenjins (Loshin 2013).

Enkele grootdata-uitdagings

Die volume, verskeidenheid en spoed waarteen grootdata groei, skep heelwat uitdagings. Een van die uitdagings is om die heterogone aard van ongestruktureerde data te hanteer, aangesien masjien-ontledingalgoritmes homogene

data verwag. Organisasies sal dus verskillende maniere benodig om die data 'gestruktureerd' te kry voordat dit vir ontleding gebruik kan word (Jagadish *et al.* 2014). Ander uitdagings is die feit dat grootdatabronne dikwels gekenmerk word deur foute, onbekendheid en onvoltooide data. Alhoewel die groot volume en oorbodigheid van grootdatastelle hierdie probleem gedeeltelik hanteer, is meer ontwikkeling nodig om rekenaarprogrammatuur te verbeter wat navrae kan hanteer op, byvoorbeeld, waarskynlikheidsdata (Jagadish *et al.* 2014). 'n Ander interessante uitdaging met grootdata is dat datavolumes tans vinniger toeneem as wat die sentrale verwerkingseenheid (SVE) klokspoed verbeter (Jagadish *et al*. 2014). Waar voorheen staatgemaak is op paralleldataverwerking oor verskeie nodes, word daar nou meer aandag geskenk aan parallellisasie op een node met 'n groter hoeveelheid kerns. Alternatiewe soos om, onder meer, van wolkgebaseerde oplossings gebruik te maak, vereis egter meer navorsing (Jagadish *et al.* 2014). Volgens Jagadish *et al.* (2014) en Hu *et al*. (2014) is privaatheid, sekuriteit en eienaarskap van data belangrike aspekte van grootdata wat meer aandag moet geniet, aangesien organisasies nou toegang het tot meer datastelle wat verband hou met hulle ondernemings as voorheen, en gebruikers onwetend data aanstuur. Phillips-Wren *et al.* (2015) voorsien 'n breedvoerige opsomming van navorsingsvrae en onopgeloste probleme wat van groot waarde vir navorsers en akademici kan wees.

Grootdata in hoër onderwys

Ten einde die waarde en geleenthede van grootdata in hoër onderwys te verstaan, word die verskillende bronne en tipes bespreek, gevolg deur geleenthede en laastens die stand van grootdata- en datawetenskaponderrig aan Suid-Afrikaanse universiteite.

Bronne en tipes grootdata in hoër onderwys

Leerbestuurstelsels (LBS), soos Blackboard of Moodle, versamel reeds volumes groeiende data van studente se akademiese prestasie, stelselinteraksie, sowel as persoonlike en demografiese inligting (Romero & Ventura 2010). Hierdie data kan gebruik word om die profiel saam te stel van 'n student wat moontlik sou belangstel om by 'n betrokke universiteit te studeer. Dit kan weer aangevul word deur sentimentdata vanaf Twitter om 'n universiteit se sentimenttelling te bepaal.

Grafiekanalise kan ook gebruik word om 'n bestaande student se sosiale netwerk te analiseer om te bepaal watter vriende ook sou belangstel om by 'n betrokke universiteit te studeer (Schmarzo 2014). Grootdatastelle vanaf sosiale media, soos Linkedin.com en Facebook.com, kan universiteite ook in staat stel om studente se gedrag beter te verstaan, asook met wie hulle sou saamwerk in 'n groep (Daniel & Butson 2014).

Ander bronne van grootdata van hoër onderwysinstellings sluit in kampussensordata sowel as data oor webblad-soekpatrone van 'n LBS. Beskrywende analise kan op hierdie datastelle toegepas word om ondersoek in te stel na hoe gereeld 'n bladsy besoek word en kan dit met die sukseskoers van 'n kursus vergelyk word om te bepaal watter studente met 'n module gesukkel het (Daniel 2015). Nog tipes grootdatabronne sluit instelling-bemarkingsdata sowel as finansiële vooruitskattingdata in, wat bestuur by universiteite in staat sou stel om proaktiewe besluite te neem in plaas van reaktief op te tree (Oracle 2015). Met inagneming van hierdie verskillende tipes grootdatabronne, is dit onafwendbaar dat universiteite in die nabye toekoms belangrike besluite ten opsigte van hul bestaande databasisstelsels sal moet neem.

Grootdatageleenthede in hoër onderwys

Een van die grootste geleenthede tans in hoër onderwys is om gekwalifiseerde studente te lewer wat in staat is om grootdatastelle te kan verwerk. In die moderne werksomgewing, wat die akademie sowel as industrie insluit, word daar dikwels van hierdie gekwalifiseerde studente verwag om as datawetenskaplikes te funksioneer (Davenport & Patil 2012). Dit vereis dat hulle op 'n wetenskaplike manier kennis en insig vanuit groot volumes gestruktureerde en ongestruktureerde data moet kan ontgin om datagedrewe besluitneming in 'n organisasie te verbeter (Dhar 2013; Provost & Fawcett 2013).

Datawetenskaplikes moet dikwels verskillende databronne gebruik om vrae soos die volgende te beantwoord: 'Wat het gebeur?', 'Waarom het dit gebeur?', 'Wat gaan gebeur?' en 'Wat is die beste wat kan gebeur?' (Van der Aalst 2014). Om hierdie vrae te kan beantwoord, word daar van datawetenskaplikes verwag om hulle sake-, data- en analitiese vaardighede te gebruik om groot ongestruktureerde datastelle te ondersoek (Sharda *et al.* 2014). Grootdatategnologie en datawetenskaptegnieke gebaseer op grootdata speel dus 'n belangrike rol om datawetenskaplikes in staat te stel om dit te vermag (Provost & Fawcett 2013). Ten einde 'n suksesvolle datawetenskaplike te word, moet datawetenskaplikes verstaan hoe grootdata geberg word (RDBS, Hadoop, Hive, NoSQL), bronkodes kan skryf (Java, C, Python), skriptale kan gebruik (Unix Bash, Perl), data kan opsoek (SQL, Hive, Pig), data kan ontleed (R, MatLab, SAS) en hulle bevindinge kan kommunikeer aan topbestuur (Watson 2014; Sharda *et al.* 2014:597; Dhar 2013). Datawetenskaplikes maak van datavisualisasie- gereedskap soos Gephi, Tableau en ggplot2 gebruik in die plek van verslae om hulle bevindinge voor te stel en aan bestuur te kommunikeer (Minelli *et al.* 2013:118; Manoochehri 2014:90). Ander vaardighede en kennis wat datawetenskaplikes benodig, sluit in teksprosessering, teksontginning en kennis oor 'n uitlegtaal soos XML (Dhar 2013). Laastens moet datawetenskaplikes data-analities kan dink (Provost & Fawcett 2013) en gaan hulle dus kennis benodig oor masjienleer, basiese statistiek soos Bayesiaanse statistiek, hipotesetoetsing en multiveranderlike ontleding, sowel as kennis oor korrelasies en die modellering van data (Dhar 2013).

Daar word gereken dat die Verenigde State van Amerika (VSA) alleen teen die jaar 2018 tussen 140,000 en 190,000 professionele mense met indringende analitiese vermoëns in grootdata gaan benodig (Berman 2013). Wixom *et al.* (2014) voer aan dat universiteite genoodsaak gaan word om bestaande grade in sake-intelligensie of sake-ontleding uit te brei en voorsiening te maak vir datawetenskap, eerder as om nuwe grade te skep wat spesifiek op datawetenskap fokus. Die Universiteit van Noord-Carolina (2016), wat gereken word as die leier op die veld van data-ontleding en datawetenskap (Davenport 2014), het onlangs 'n lys gepubliseer van meer as 80 voorgraadse kwalifikasies in sake-intelligensie, data-ontleding of datawetenskap wat by universiteite in die VSA aangebied word. Waar volledige datawetenskapprogramme aangebied is, was dit hoofsaaklik op meestersvlak gewees (Davenport 2014).

In Suid-Afrika het die aanvraag na datawetenskaplikes oor die laaste vyf jaar aansienlik toegeneem (Taylor 2015). In antwoord daarop, het die Departement van Wetenskap en Tegnologie (2015) die Inter-Universitêre Instituut vir Data-intensiewe Astronomie (IDIA) geloods om aan die behoeftes van onder meer die Suid-Afrikaanse SKA-projek (Square Kilometre Array) te voorsien. Die SKA-projek word gesien as een van twee wêreldwye projekte wat grootdata binne astronomie gaan dryf (Universiteit van Kaapstad 2013). Daar word voorsien dat die IDIA-projek kenners uit die veld van astronomie, rekenaarwetenskap en statistiek sal saambring om die kapasiteit van datawetenskap te ontwikkel, nie net vir die SKA-projek nie, maar ook ander velde in Suid-Afrika (Taylor 2015). Gedurende die afgelope paar jaar is studies ook uitgevoer oor die gebruik van grootdata in Suid-Afrika (Malaka & Brown 2015; Nartey 2015; Ridge *et al.* 2015). In hierdie studies het die tekort aan vaardighede en grootdata-strategie, asook tegnologiese uitdagings konstant na vore getree as die belangrikste grootdata-uitdagings. Ten tye van hierdie studie was dit nie duidelik in hoe 'n mate Suid-Afrikaanse universiteite studente akademies voorberei om hierdie uitdagings die hoof te bied nie. Die volgende afdeling gee terugvoer hieroor.

Stand van datawetenskap en grootdata by Suid-Afrikaanse universiteite

Die studie het ondersoek ingestel na die stand van grootdata- en datawetenskaponderrig aan Suid-Afrikaanse universiteite. Data is deur middel van internetsoektogte op Google.com versamel. Data is ook vanaf die Suid-Afrikaanse Kwalifikasie-Owerheid (SAQA) se databasis verkry. Al die data is ingesamel tussen Januarie en Maart 2016.

Die navorsingsvraag vir hierdie studie is: *Watter universiteite in Suid-Afrika bied modules in datawetenskap en grootdata aan en wat behels die inhoud?*

Ten opsigte van grootdata, is die databasis van die Suid-Afrikaanse Kwalifikasie-Owerheid (SAQA) geraadpleeg ten opsigte van kwalifikasie- en eenheidstandaarde waar die titel '*big data*' voorkom. Die soektog het geen resultate opgelewer nie en is uitgebrei met woordsoektogte wat die terme '*nosql*', '*hdfs*', '*mapreduce*' en '*hadoop*' ingesluit het. Die soektog het weereens geen resultate opgelewer nie en daar is besluit om gebruik te maak van internetsoektogte. 'n Lys van topuniversiteite in Suid-Afrika is volgens die *University Ranking* by *Academic Perfomance* (URAP 2016) geïdentifiseer. Hierdie universiteitelys is gebruik in uitgebreide internetsoektogte om te bepaal watter van die universiteite gevorderde databasismodules op nagraadse vlak aanbied. Dié modules het ingesluit datapakhuise (DP), sake-intelligensie (BI) en grootdata (GD). Bronne wat geraadpleeg is, sluit in jaarboeke sowel as modulehandleidings. Uit die lys universiteite, het net vyf universiteite 'n gevorderde databasiskursus in datapakhuise en/of sake-intelligensie aangebied. Van die vyf universiteite, het net twee grootdata-inhoud ingesluit in 'n module. Tabel 1 lys die universiteite met 'n DP/BI module-aanwyser sowel as 'n GD-aanwyser. Die aanwyser is as 'Ja' gemerk indien die universiteit die inhoud aanbied en 'Nee' andersins.

Die inhoud van Noordwes-Universiteit (2015) se twee nagraadse, gevorderde databasisstelsels-modules (ITRI611 en ITRI621) het gehandel oor datapakhuise en sake-intelligensie. Studente moes as 'n praktiese projek 'n datapakhuis en OLAP-kubus bou en inhandig vir assessering, tesame met genoegsame dokumentasie (Goede 2014). Geen grootdata-teorie of -tegnologie is in die twee modules aangebied nie. Die inhoud van die Universiteit van Pretoria (2015) se gevorderde databasisstelsel-module (INF785) het gehandel oor gevorderde databasisontwerp, databasisargitektuur, datapakhuise en datamarte. Geen grootdata-teorie of -tegnologie is aangebied in die module nie. Die Nelson Mandela Metropool-Universiteit (2015) het ook soos Noordwes-Universiteit, twee nagraadse gevorderde databasisstelsels-modules aangebied. Die inhoud van die twee modules, WRDW411 en WRBI411, het datapakhuise, data-ontginning, sake-intelligensie en OLAP behandel. Geen grootdata-teorie of -tegnologie is aangebied in die twee modules nie.

Die Universiteit van Kaapstad (2015) se nagraadse, gevorderde databasisstelsel-module (DBS) het inhoud oor verspreide databasis, datapakhuise, data-ontginning, asook grootdata en NoSQL ingesluit. Die doel van die module was om studente bloot te stel aan nuwe ontwikkelings in

TABEL 1: Universiteite in Suid-Afrika wat DP/BI en/of GD.

Rang in Afrika	Universiteit	DP/BI	GD
1	Universiteit van Kaapstad	Ja	Ja
6	Universiteit van Pretoria	Ja	Nee
11	Noordwes-Universiteit	Ja	Nee
18	Universiteit van die Vrystaat	Ja	Ja
32	Nelson Mandela Metropool-Universiteit	Ja	Nee

DP/BI, datapakhuise/sake-intelligensie; GD, grootdata.

Nota: Datapakhuise, sake-intelligensie en grootdata is opgestel deur ondersoek in te stel of die instansies datapakhuis, sake-intelligensie en grootdata aanbied.

databasisstelsels en om nuwe tegnologie wat daarmee gepaardgaan, te bestudeer. Uit die modulehandleiding was dit nie duidelik hoe indringend grootdata en NoSQL behandel word nie. Die Universiteit van die Vrystaat (2015) het drie nagraadse, gevorderde databasisstelsel-modules aangebied, naamlik Oracle- databasisadministrasie (CSID6843), datapakhuise (CSID6853) en sake-intelligensie (CSID6823). Grootdata-teorie en toepassing word as deel van die datapakhuis-module aangebied. Die inhoud het tegnologieë soos Hadoop, MapReduce, HDFS en Hive ingesluit en studente se vaardighede is deur teoretiese en praktiese opdragte geassesseer. Die gebruik van datavisualiseringstegnologie, meer spesifiek Tableau, word as deel van die nagraadse sake-intelligensie-module aangebied.

Die studie het vervolgens ondersoek ingestel na datawetenskaponderrig aan Suid-Afrikaanse universiteite. Die Suid-Afrikaanse Kwalifikasie-Owerheid (SAQA) se databasis is weereens geraadpleeg met die soekterm '*data science*' (SAQA 2015). Die soektog is aangevul deur uitgebreide internetsoektogte oor datawetenskapkwalifikasies aan Suid-Afrikaanse universiteite. Tabel 2 stel die resultaat van die soektogte voor.

Die kwalifikasies se inhoud is vergelyk met datawetenskapvaardighede uit die literatuur. Sol Plaatje-Universiteit (2016) se kwalifikasie fokus hoofsaaklik op Wiskunde en Rekenaarwetenskap. Ten opsigte van die vaardighede wat benodig word, is daar net vermelding gemaak van Microsoft SQL wat as bediener gebruik word om data te stoor, te bestuur en navrae met SQL te implementeer. Dit dek net 'n gedeelte van hoe grootdata geberg word, en meer spesifiek gestruktureerde data wat RDBS-tegnologie, soos Microsoft SQL gebruik. Noordwes-Universiteit (2016) se kwalifikasie dek bedryfsanalise en data-ontginning in die hoofstroom en maak vermelding van SAS Enterprise Miner as data-ontginningsgereedskap. Hierdie gereedskap sluit aan by die data-ontledingsvaardighede wat datawetenskaplikes benodig (Watson & Marjanovic 2014).

Opsomming en toekomstige werk

Die snelle groei van grootdata en gepaardgaande tegnologie bied astronomiese geleenthede vir die industrie sowel as die akademie in Suid-Afrika. Die doel van die artikel is om 'n breedvoerige oorsig te gee oor grootdata, grootdatastelsels, relevante oopbronprogrammatuur wat met grootdatastelsels geassosieer word, sowel as grootdata-onderrig. Grootdata- en datawetenskaponderrig is 'n belangrike en 'n relevante onderwerp vir Suid-Afrika, aangesien die vraag na datawetenskaplikes die afgelope vyf jaar aansienlik toegeneem het.

TABEL 2: Datawetenskapkwalifikasies.

Universiteite	Graad
Sol Plaatje-Universiteit	*B.Sc. in Data Science*
Noordwes-Universiteit	*M.Sc. in Business Mathematics and Informatics*

Die artikel rapporteer dat grootdata- en datawetenskaponderrig aan Suid-Afrikaanse universiteite wel aandag geniet. Daar word grootdatavaardighede aan nagraadse studente oorgedra, wat insluit hoe data geberg word (RDBS, HDFS, HBase), data verwerk word (Hadoop, MapReduce, Spark), data opgesoek word (Hive, SQL) en data gevisualiseer word (Tableau), maar op 'n beperkte skaal. Dit is egter nog nie seker of bestaande Suid-Afrikaanse akademiese programme genoegsame blootstelling aan nagraadse studente gee ten opsigte van die wetenskap, sowel as die grootdatategnologiese komponent van datawetenskap nie. Die volgende inhoud word aanbeveel as deel van toekomstige blootstelling aan en onderrig in grootdatategnologie:

- Bedryfstelsel: UNIX-verwante bedryfstelsel (Bash-tekste)
- Grootdataberging: NoSQL, Hadoop, HDFS
- Grootdata-programmeringmodel: MapReduce, Spark
- Grootdataprogrammering: Python, Perl of Java
- Grootdata-ontleding: Hive, SQL
- Grootdata-ontginning: Mahout, R, SAS Enterprise Miner, IBM SPSS Modeler
- Grootdatavisualisering: Tableau, Gephi of ggplot2.

Grootdata-onderrig gaan in die nabye toekoms 'n belangrike rol speel om in die vinnig toenemende tekort aan datawetenskaplikes te voorsien. Hierdie datawetenskaplikes se toekomstige bydraes in die sake- en wetenskapwêreld moet nie onderskat word nie, en het die potensiaal om aan Suid-Afrika 'n regmatige plek in die wêreld van grootdata te gee. Meer navorsing word vereis om te bepaal wat Suid-Afrika se werklike behoefte aan datawetenskaplikes met toepaslike grootdata-onderrig is. Verdere navorsing in terme van 'n ondersoek om vas te stel of die grootdatategnologie wel, soos voorgestel, deur organisasies in Suid-Afrika gebruik word, is egter nodig. Hierdie navorsing kan 'n waardevolle en belangrike bydrae lewer ten einde die akademiese opleiding van datawetenskaplikes in Suid-Afrika te bevorder.

Erkenning

Mededingende belange

Die outeur verklaar hiermee dat hy geen finansiële of persoonlike verbintenisse het wat hom nadelig of voordelig beïnvloed het in die skryf van hierdie artikel nie.

Literatuurverwysings

Apache Spark, 'Apache Spark documentation and open source release', viewed on 21 June 2016, from http://spark.apache.org/

Assunção, M.D., Calheiros, R.N., Bianchi, S., Netto, M.A.S. & Buyya, R., 2015. 'Big data computing and clouds: Trends and future directions', *Journal of Parallel and Distributed Computing*, 79, pp. 3–15. http://dx.doi.org/10.1016/j.jpdc.2014.08.003

Berman, J., 2013, *Principles of big data: Preparing, sharing, and analyzing complex information*, Elsevier, Morgan Kaufmann, Amsterdam.

Celko, J., 2014, *Complete guide to NoSQL*, Morgan Kaufmann, Burlington, MA.

Chang, F., Dean, J., Ghemawat, S., Hsieh, W.C., Wallach, D.A., Burrows, M., Chandra, *et al.*, 2006, 'Bigtable: A distributed storage system for structured data', *7th Symposium on Operating Systems Design and Implementation (OSDI '06)*, November 6–8, Seattle, WA, USA, pp. 205–218.

Chen, M., Mao, S., Zhang, Y. & Leung, V. C., 2014, *Big data: Related technologies, challenges and future prospect*.

Daniel, B., 2015, 'Big data and analytics in higher education: Opportunities and challenges', *British Journal of Educational Technology*, 46(5), pp. 904–920. http://dx.doi.org/10.1111/bjet.12230

Daniel, B. & Butson, R., 2014, 'Foundations of big data and analytics in higher education, in *International Conference on Analytics Driven Solutions: ICAS2014*.

Databricks, 2016, Apache Spark Ecosystem, viewed on 21 June 2016, from https://databricks.com/spark/about

Davenport, T.H., 2014, *Big data at work*, Harvard Business School Publishing Corporation, Boston, Massachusetts.

Davenport, T.H. & Patil, D.J., 2012, 'Data scientist: The sexiest job of the 21st century', *Harvard business review*, 90. http://dx.doi.org/10.15358/9783800648153

Department of Science and Technology, 2015, minister Naledi Pandor launched the Inter-University Institute for Data Intensive Astronomy (IDIA) | South African Government, viewed on 07 March 7 2016, from: http://www.gov.za/speeches/minister-naledi-pandor-launch-inter-university-institute-data-intensive-astronomy-idia-3

Dhar, V., 2013, 'Data science and prediction', *Communications of the ACM*, 56(12), pp. 64–73. http://dx.doi.org/10.1145/2500499

Eckerson, W., 2011, 'Big data analytics: Profiling the use of analytical platforms in user organizations', *TDWI*, (September), pp. 1–49, viewed on 11 November 2015, from http://tdwi.org/login/default-login.aspx

Gartner, 2016, 'What is big data? – Gartner IT Glossary – Big Data', viewed on 17 February 2016, from http://www.gartner.com/it-glossary/big-data/

Goede, R., 2014, 'An instructional design for data warehousing : Using design science research and project-based learning', *International Journal of Technology and Inclusive Education*, 1(3), pp. 493–499.

Hadoop, n.d., 'Hadoop documentation and open source release', viewed on 08 March 2016, from: http://hadoop.apache.org/

Hu, H., Wen, Y., Chua, T. & Li, X. , 2014, 'Toward scalable systems for big data analytics: A technology tutorial', *Access, IEEE*, vol. 2, pp. 652–687. http://dx.doi.org/10.1109/ACCESS.2014.2332453

IBM, n.d., 'What is big data?' viewed on 06 January 2015, from http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

Intel, 2013, 'Extract, transform, and load big data with Apache Hadoop', p. 9, viewed on 03 February 2016, from https://software.intel.com/en-us/articles/extract-transform-and-load-big-data-with-apache-hadoop

Jacobs, A., 2009, 'The pathologies of big data, *Communications of the ACM*, 52(8), pp. 36–44. http://dx.doi.org/10.1145/1536616.1536632

Jagadish, H.V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R. & Shahabi, C., 2014, 'Big data and its technical challenges', *Communications of the ACM*, 57(7), pp. 86–94. http://dx.doi.org/10.1145/2611567

Kim, B.G., Trimi, S. & Chung, J., 2014, 'Big-Data applications in the government sector', *Communications of the ACM*, 57(3), pp. 78–85. http://dx.doi.org/10.1145/2500873

Kim, W., Jeong, O.-R. & Kim, C., 2014, 'A holistic view of big data', *International Journal of Data Warehousing and Mining*, 10(3), pp. 59–69. http://dx.doi.org/10.4018/ijdwm.2014070104

Krishnan, K., 2013, *Data warehousing in the age of big data*, Elsevier, Morgan Kaufmann, Amsterdam.

Landset, S., Khoshgoftaar, T. M., Richter, A. N. & Hasanin, T., 2015, 'A survey of open source tools for machine learning with big data in the Hadoop ecosystem', *Journal of Big Data*, 2(1), p. 24. http://dx.doi.org/10.1186/s40537-015-0032-1

Lee, K.H., Lee, Y.J., Choi, H., Chung, Y.D. & Moon, B., 2011, 'Parallel data processing with MapReduce: a survey', *ACM SIGMOD Record*, 40(4), p. 11. http://dx.doi.org/10.1145/2094114.2094118

Leverenz, L., 2015, 'Tutorial – Apache Hive – Apache Software Foundation', viewed on 09 February 2016, from https://cwiki.apache.org/confluence/display/Hive/Tutorial#Tutorial-LoadingData

Loshin, D., 2013, *Big data analytics: From strategic planning to enterprise integration with tools, techniques, NoSQL, and graph*, Elsevier, United States of America.

Malaka, I. & Brown, I., 2015, 'Challenges to the organisational adoption of big data analytics', in *Proceedings of the 2015 Annual Research Conference on South African Institute of Computer Scientists and Information Technologists - SAICSIT '15*, ACM Press, pp. 1–9, New York, NY. http://dx.doi.org/10.1145/2815782.2815793

Manoochehri, M., 2014, *Data just right*, Pearson, Crawfordsville, Indiana.

Minelli, M., Chambers, M. & Dhiraj, A., 2013, *Big data, big analytics: Emerging business intelligence and analytic trends for today's businesses*, John Wiley & Sons, Hoboken, NJ. http://dx.doi.org/10.1002/9781118562260

Nartey, C., 2015, 'Entrepreneurship and innovation management big data and start-up businesses', *Entrepreneurship and Innovation Management Journal*, 3(3), pp. 155–163.

Nelson Mandela Metropole University, 2015, 'Postgraduate Modules', viewed on 09 February 2016, from http://cs.nmmu.ac.za/Postgraduate-Information/Postgraduate-Modules

Northwest University 2016, 'BMI brochure', viewed on 16 February 2016, from http://www.nwu.ac.za/sites/www.nwu.ac.za/files/files/p-bmi/documents/Bemarkingsmateriaal/Marketing brochure for prospective students.pdf

Northwest University, 2015, 'Information Sheet: Honours students 2015' viewed on 09 February 2016, from http://www.nwu.ac.za/sites/www.nwu.ac.za/files/files/v-fesit-sit/it courses/Final InformationSheet 2015.pdf

Oracle, 2015, Improving higher education performance with big data architect' s guide and reference architecture introduction, viewed on 29 March 2016, from http://www.oracle.com/us/technologies/big-data/big-data-education-2511586.pdf

Phillips-Wren, G., Iyer, L. S., Kulkarni, U. & Ariyachandra, T., 2015, 'Business analytics in the context of big data : A roadmap for research', *Communications of the Association for Information Systems*, 37, p. Article 23.

Provost, F. & Fawcett, T., 2013, 'Data science and its relationship to big data and data-driven decision making', *Data Science and Big Data*, 1(1).

Rabkin, A. & Katz, R., 2010, 'Chukwa: A system for reliable large-scale log collection', *Proceeding LISA'10 Proceedings of the 24th International Conference on Large Installation System Administration*, pp. 1–15.

Rahman, N. & Iverson, S., 2015, 'Big data business intelligence in bank risk analysis', *International Journal of Business Intelligence Research*, 6(2), pp. 55–77, viewed on 30 March 2016, from http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/IJBIR.2015070104

Ridge, M., Kevin, A.J. & Brian, D., 2015, 'The use of big data analytics in the retail industries in South Africa', *African Journal of Business Management*, 9(19), pp. 688–703. http://dx.doi.org/10.5897/AJBM2015.7827

Romero, C. & Ventura, S., 2010, 'Educational data mining: A review of the state of the art', *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6), pp. 601–618. http://dx.doi.org/10.1109/TSMCC.2010.2053532

Russom, P., 2015, 'Can Hadoop replace a data warehouse?', viewed on 22 June 2016, from https://tdwi.org/articles/2015/01/27/hadoop-replace-data-warehouse.aspx

Sabherwal, R. & Becerra-Fernandez, I., 2011, *Business intelligence: Practices, technologies and management*, Wiley.

SAQA, 2015, South African Qualifications Authority, viewed on 02 November 2015, from http://www.saqa.org.za/

Schmarzo, B., 2014, 'What universities can learn from big data – higher education analytics', viewed 18 March 2016 from https://infocus.emc.com/william_schmarzo/what-universities-can-learn-from-big-data-higher-education-analytics/

Sharda, R., Delen, D. & Turban, E., 2014, '*Pearson Education – Business intelligence and analytics: Systems for decision Support, Global Edition* 10th ed., Pearson Education, Boston.

Sol Plaatje University, 2016, 'Bachelor of Science in Data Science', viewed on 16 February 2016, from http://allqs.saqa.org.za/showQualification.php?id=96105

Taylor, R., 2015, 'Data Science is the way of the future', viewed on 07 March 2016, from http://www.up.ac.za/en/news/post_2175699-data-science-is-the-way-of-the-future

University of Cape Town, 2015, 'Honours Handbook 2015', viewed on 09 February 2016, from http://www.cs.uct.ac.za/honours/Honours_Handbook_2015.pdf

University of Cape Town, 2013, 'The big data revolution', viewed on 01 March 2016, from http://www.uct.ac.za/dailynews/?id=8767

University of the Free State, 2015, 'Rule Book 2015', viewed on 09 February 2016, from http://apps.ufs.ac.za/dl/yearbooks/237_yearbook_eng.pdf

University of North Carolina, 2016, 'Degree programs in analytics and data science', viewed on 04 March 2016, from http://analytics.ncsu.edu/?page_id=4184

University of Pretoria, 2015, 'Yearbooks 2015', viewed on 09 February 2016, from http://www.up.ac.za/yearbooks/modules/view/INF785

URAP, 2016, University Ranking by Academic Performance, viewed on 05 February 2016, from http://www.urapcenter.org/2015/region.php?rcode=AF

Van der Aalst, W.M., 2014, 'Data scientist: The engineer of the future', in *Enterprise Interoperability VI*. Springer International Publishing, pp. 13–26. http://dx.doi.org/10.1007/978-3-319-04948-9_2

Vaisman, A. & Zimanyi, E., 2014, *Data warehouse systems design and implementation*, Springer-Verlag, Heidelberg.

Vavilapalli, V.K., Murthy, A.C., Douglas, C., Agarwal, S., Konar, M., Evans, R., Graves, T., Lowe, J., Shah, H., Seth, S. and Saha, B., 2013, 'Apache Hadoop YARN', in *Proceedings of the 4th Annual Symposium on Cloud Computing - SOCC '13,* ACM Press, pp. 1–16, New York, NY. http://dx.doi.org/10.1145/2523616.2523633

Wamba, S.F., Akter, S., Edwards, A., Chopin, G. and Gnanzou, D., 2015. 'How "big data" can make big impact: Findings from a systematic review and a longitudinal case study', *International Journal of Production Economics*.

Warden, P., 2011, *Big data glossary*, O'Reilly Media, Sebastopol, CA.

Watson, H.J., 2014, 'Tutorial: Big data analytics: Concepts, technologies, and applications', *Communications of the Association for Information Systems*, 34(1), p. 64.

Watson, H.J. & Marjanovic, O., 2014, 'Big data : The fourth data management generation', *Business Intelligence Journal*, 18(3), pp. 4–9.

White, T., 2015, *Hadoop: The definitive guide*, 4th ed., O'Reilly Media, Sebastopol, CA.

Wixom, B., Ariyachandra, T., Douglas, D., Goul, M., Gupta, B., Iyer, L., Kulkarni, U., Mooney, J.G., Phillips-Wren, G. and Turetken, O., 2014, 'The current state of business intelligence in academia: The arrival of big data', *Communications of the Association for Information Systems*, 34, pp. 1–13.

Yin, S. & Kaynak, O., 2015, 'Big data for modern industry: Challenges and trends', *Proceedings of the IEEE*, 103(2), pp. 143–146. http://dx.doi.org/10.1109/JPROC.2015.2388958

Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S. and Stoica, I., 2010, 'Spark: Cluster computing with working sets', *Proceedings of the 2nd USENIX conference on hot topics in cloud computing*, p. 10.